

ECOLOGY

Building a genome-based understanding of bacterial pH preferences

Josep Ramoneda^{1*}, Elias Stallard-Olivera^{1,2}, Michael Hoffert^{1,2}, Claire C. Winfrey^{1,2}, Masumi Stadler³, Juan Pablo Niño-García^{3,4}, Noah Fierer^{1,2*}

The environmental preferences of many microbes remain undetermined. This is the case for bacterial pH preferences, which can be difficult to predict a priori despite the importance of pH as a factor structuring bacterial communities in many systems. We compiled data on bacterial distributions from five datasets spanning pH gradients in soil and freshwater systems (1470 samples), quantified the pH preferences of bacterial taxa across these datasets, and compiled genomic data from representative bacterial taxa. While taxonomic and phylogenetic information were generally poor predictors of bacterial pH preferences, we identified genes consistently associated with pH preference across environments. We then developed and validated a machine learning model to estimate bacterial pH preferences from genomic information alone, a model that could aid in the selection of microbial inoculants, improve species distribution models, or help design effective cultivation strategies. More generally, we demonstrate the value of combining biogeographic and genomic data to infer and predict the environmental preferences of diverse bacterial taxa.

INTRODUCTION

Predicting the environmental preferences of organisms is an important goal in ecology. If we know the conditions under which a given taxon can thrive, then we can better predict biogeographical distributions (1), guide ecological restoration efforts (2), design effective probiotics (3), and understand taxon-specific responses to global change factors (4). Unfortunately, the environmental preferences of most microbial taxa and the genomic attributes associated with those preferences often remain undetermined (5, 6). One reason for this is that most microorganisms, particularly those found in nonhost-associated environments, can be difficult to cultivate in vitro (7), making it difficult to measure environmental preferences directly. Even for those taxa that can be cultivated, quantifying how microbial growth rates vary across broad environmental gradients can be time consuming and the environmental gradients created in vitro may not necessarily mimic those found in situ. However, when direct information on environmental preferences can be collected, such information can be very useful for predicting microbial distributions and functions across space and time [e.g., (8–11)].

Even without the direct measurement of environmental preferences, it is feasible to infer some environmental preferences from genomic information (12). We can leverage the information contained in curated genomic databases, which can include both cultivated and uncultivated microbial taxa (13), to infer the specific environmental preferences of uncharacterized microorganisms (14). For example, genomic information from isolates whose environmental preferences have been measured in vitro has been used to infer the preferences of bacteria across gradients in oxygen (15) and temperature (12). Using genomic information to determine the

environmental preferences of microbial taxa can have important ramifications. For example, we could improve our ability to predict community assembly across different environmental gradients, identify the conditions under which specific taxa can thrive, and better optimize medium conditions to improve the cultivability of fastidious taxa. However, genome-based inferences of environmental preferences can be difficult to validate (especially for uncultivated taxa), and we do not always know which genes or other genomic attributes are associated with adaptations to specific environmental conditions of interest.

Consider microbial preferences for specific pH conditions. To our knowledge, it is not now possible to predict bacterial pH preferences from genomic information alone, although we know that pH is often a key factor determining the niche space occupied by microorganisms. The distributions of specific bacterial taxa and the overall composition of bacterial communities are often strongly associated with gradients in pH, as has been observed in a wide range of environments including soils (16–19), freshwater (20–22), and geothermal systems (23). Despite the importance of this environmental factor, the pH preferences of most bacterial taxa remain undetermined, although we know that bacterial pH preferences can vary widely (24). This knowledge gap is particularly evident in nonhost-associated systems that are often dominated by taxa that are difficult to cultivate and study in vitro. Even across cultivated taxa, pH tolerances and pH optima for growth are rarely determined experimentally, and most cultivation media likely select for taxa that grow at near-neutral pH conditions (25). We note that “pH preference” across a given gradient is akin to the “realized niche” of a population (as opposed to the “fundamental niche”), where pH preference is the pH at which an organism achieves maximal relative abundances in nature (17). This relative abundance is determined by the metabolically optimal pH for growth as well as other biotic or abiotic constraints. For example, a given taxon may grow optimally at around pH 7 under controlled conditions, but its pH preference could be lower if its abundance in

Copyright © 2023 The Authors, some rights reserved; exclusive licensee American Association for the Advancement of Science. No claim to original U.S. Government Works. Distributed under a Creative Commons Attribution NonCommercial License 4.0 (CC BY-NC).

¹Cooperative Institute for Research in Environmental Sciences, University of Colorado, Boulder, CO, USA. ²Department of Ecology and Evolutionary Biology, University of Colorado, Boulder, CO, USA. ³Département des Sciences Biologiques, Université du Québec à Montréal, Case Postale 8888, Succursale Centre-Ville, Montréal, QC H3C 3P8, Canada. ⁴Escuela de Microbiología, Universidad de Antioquia, Ciudad Universitaria Calle 67 No 12 53-108, Medellín, Colombia.

*Corresponding author. Email: ramoneda.massague@gmail.com (J.R.); noah.fierer@colorado.edu (N.F.)

a given environment is maximized at pH 6 due to its interplay with other factors.

In the absence of whole genome sequence data, one might still be able to predict bacterial pH preferences from taxonomic and phylogenetic information alone. Shifts in bacterial community composition across pH gradients can be evident at broad levels of taxonomic resolution (26, 27), suggesting some degree of conservatism in the traits related to pH adaptation. However, closely related taxa can have very distinct pH preferences (28), such as bacteria within the phylum Acidobacteria which tend to be more abundant in lower pH soils (27), a pattern that does not necessarily hold for subgroups within the phylum (29, 30). Thus, it is not clear whether taxonomy and phylogenetic information is sufficient to predict bacterial pH preferences. Previous studies have focused on identifying common adaptations to changes in pH conditions across different taxonomic groups and the genes or transcripts associated with those adaptations (28, 31–33). This generates the expectation that the presence or absence of specific functional genes can be used to predict bacterial pH preferences. An approach that integrates biogeographic distributions of multiple individual taxa with genomic information across environmental gradients can address whether taxonomic, phylogenetic, or genomic information can be used to predict microbial environmental preferences.

We set out to determine whether bacterial pH preferences are predictable. In other words, we asked whether we could use taxonomic, phylogenetic, or genomic information to predict where along a pH gradient a taxon is most likely to achieve its highest relative abundance (which we define here as its “pH preference”), searching for patterns that are generalizable across distinct environment types. To do so, we used information on bacterial distributions across five independent datasets that span large pH gradients in soil and freshwater systems to infer the putative pH preferences of the bacterial taxa found in these environments. We then used this information to assess the degree to which bacterial pH preferences are phylogenetically and taxonomically conserved. To determine the genomic features associated with adaptations to pH, we analyzed representative genomes from taxa with varying pH preferences, as inferred from our cultivation-independent analyses, and identified genes that are consistently associated with differences in bacterial pH preference across environments. Last, we developed and validated a machine learning model that enables the accurate identification of bacterial pH preferences from the presence or absence of 56 functional genes, making it feasible to infer pH preferences for both uncultivated and cultivated taxa. More generally, we demonstrate how our workflow (Fig. 1) can be used to investigate other bacterial environmental preferences and the genes associated with environmental adaptations while overcoming the limitations of cultivation-based experimental approaches to expand our trait-based understanding of microorganisms.

RESULTS AND DISCUSSION

Overview of the approach

We first inferred bacterial pH preferences using biogeographical information from natural pH gradients found in distinct systems (soil and freshwater environments) (Fig. 1A). We acknowledge that pH is unlikely to be the only factor influencing the distributions of bacteria in these systems and that the pH preferences of many taxa could not be inferred using this approach as there are other biotic or

abiotic factors that are of similar, or greater, importance in shaping their distributions. However, we note that this approach is similar to the approach routinely used in plant and animal systems to quantify the relationships between environmental factors and growth optima or tolerances (34). After inferring pH preferences from the biogeographical information, we then matched the 16S ribosomal RNA (rRNA) gene sequences from those taxa for which pH preferences could be inferred to representative genomes, identifying sets of genes that are consistently associated with pH preference based on their presence or absence in a given genome (Fig. 1B). These genes were then incorporated into a machine learning modeling framework (Fig. 1C) to predict pH preferences, with the approach validated using independent test sets.

The biogeographical analyses included 16S rRNA gene sequence data from a total of 795 soil samples and 675 freshwater samples spanning a pH range from 3 to 10 with a total of 250,275 amplicon sequence variants (ASVs) included in downstream analyses (fig. S1 and table S1). These data came from five independent datasets, and, in all cases, pH was an important driver of overall bacterial community composition, with Mantel ρ values ranging from 0.37 [La Romaine watershed (ROMAINE), freshwater] to 0.78 [Panama (PAN), tropical forest soils] (fig. S1). These patterns are expected considering that pH is often observed to have a strong influence on bacterial community structure in many systems (27, 35) and considering that each of these sample sets was specifically selected to span broad gradients in pH. We also note that a broad diversity of bacterial taxa was found within and across each of the five sample sets (fig. S2A), which is important as we were trying to identify patterns in pH preferences across a broad array of taxa. Using our conservative approach, we were able to estimate the pH preferences of bacterial taxa in these environmental samples for 0.5 to 4.9% of all ASVs per dataset (table S1). The analyses of the taxonomic and phylogenetic signals in bacterial pH preferences and the search for representative genomes from these taxa were ultimately based on a total of 4568 ASVs (468 to 1614 ASVs per dataset) spanning 38 bacterial phyla.

Is taxonomic and phylogenetic information good predictors of the pH preference of bacterial taxa?

Taxonomy was a poor predictor of bacterial pH preferences (Fig. 2). In almost every phylum, there were ASVs with very distinct pH preferences and there were few cases where a high proportion of ASVs from a particular phylum were found to have a similar pH preference. For example, in the ROMAINE freshwater dataset, many ASVs assigned to the phylum Acidobacteria did exhibit a general preference for acidic pH conditions, but this observation was not consistent across datasets (Fig. 2). Thus phylum-level information provides relatively little insight into bacterial pH preferences, most likely because taxa within a given phylum can have divergent pH preferences [as noted previously for different acidobacterial subdivisions (30, 36)]. We observed similar patterns at finer levels of taxonomic resolution. For example, ASVs within some of the most ubiquitous bacterial families observed across these five datasets (Xanthobacteraceae in soil and Chitinophagaceae in freshwater systems) included ASVs with inferred pH preferences ranging from 4.01 to 8.20 and from 4.63 to 8.35, respectively.

Consistent with the taxonomy-based results, we also found that bacterial pH preferences were not readily predictable from phylogenetic information. In other words, there was minimal phylogenetic

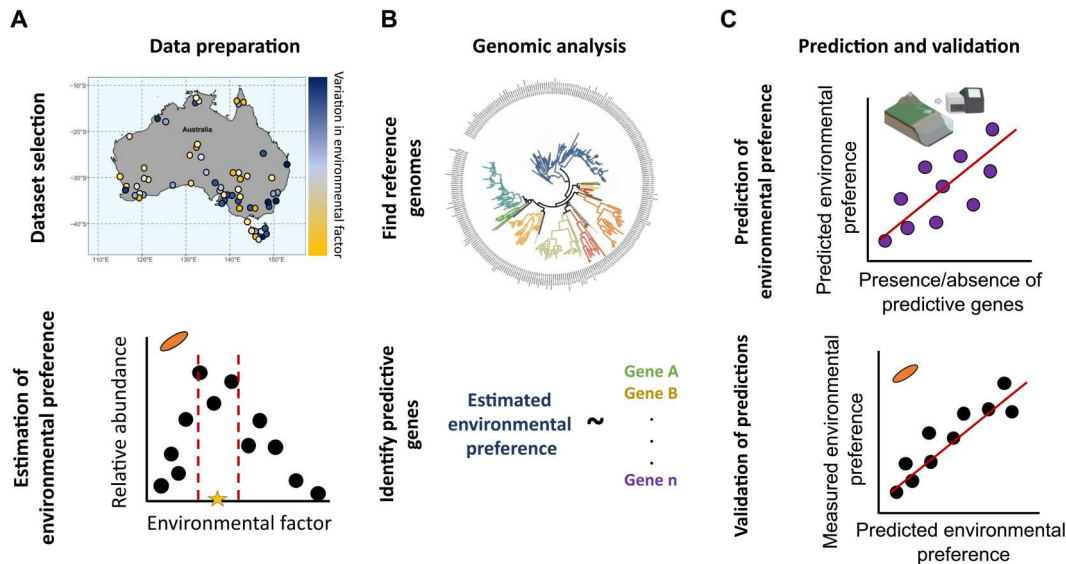


Fig. 1. Workflow to predict environmental preferences from genomic and biogeographical information. (A) Marker gene sequencing datasets that include samples spanning a broad range in the environmental factor of interest (pH in our case) are chosen and rare taxa are excluded on the basis of occurrence. The remaining taxa are then used for the estimation of the environmental conditions at which the relative abundances of individual taxa are highest using either bootstrapping or generalized additive model approaches. (B) The taxa for which an environmental preference has been statistically estimated are then queried against curated genome databases. Genomes matching the query sequence at high similarity (e.g., at 99.6% in this study) are considered representative of the taxa. (C) All genes present in those genomes are then tested for their relationship to environmental preference, and those that are significantly associated with environmental preference (e.g., sharing the same statistically significant positive or negative relationship with pH across datasets) are incorporated into a machine learning modeling framework. Predictions can be validated using an independent test set of the data, from pH preferences estimated in independent datasets, or from phenotypic assays.

conservation in pH preferences. Although we detected a significant phylogenetic signal in pH preferences for all but one dataset (Fig. 3 and fig. S3), the signal was relatively weak (Pagel's λ , 0.22 to 0.78) (fig. S3). This is qualitatively evident from the phylogenetic trees associated with each dataset, which show that even closely related taxa often had very distinct pH preferences. For example, within the Proteobacteria phylum, some clusters of ASVs with similar preference for higher or lower pH conditions were indeed observed in the phylogenies from both Australia (AUS) (soil) and ROMAINE (freshwater) datasets, but these clusters were not evident in the other datasets (Fig. 3). These observations were further confirmed by the phylogenetic correlogram analyses that show that the Moran's I autocorrelation values were low in all cases [<0.25 ; (37)], even at shallow phylogenetic depth (figs. S2B and S3). A significant Pagel's λ indicates that there is a significant phylogenetic signal of the trait by comparison to a random model of evolution, while Moran's I indicates whether this signal is clustered around a particular phylogenetic depth based on correlation. Because these indices are differently affected by sample size and trait variation, there is no consensus on thresholds that help interpret these parameters together (37). A statistically significant Pagel's λ verifies that there is phylogenetic signal, and the magnitude of Moran's I indicates how strong this signal is at a given phylogenetic depth. Our results are in line with previous work that found no phylogenetic conservation of pH preferences across both cultured bacteria (14) and uncultured bacteria (38). Neither taxonomic nor phylogenetic information is generally useful for determining pH preferences without additional information, highlighting the potential value of incorporating functional gene information to better predict bacterial pH preferences.

Associations between bacterial pH preferences and functional genes

We next analyzed representative genomes of those ASVs with inferred pH preferences. Expectedly, given the well-recognized biases in reference genome databases (39), only a relatively small fraction (6 to 24%) of those ASVs identified in our sample sets had representative genomes in Genome Taxonomy Database (GTDB) (40) and a substantial portion (~35%) of those genomes were metagenome-assembled genomes (MAGs) and single-cell-assembled genomes (SAGs) from uncultured taxa (fig. S4). The number of genome matches obtained ranged from 57 genomes in the AUS soil dataset to 293 genomes in the Carbon Biogeochemistry in Boreal Aquatic Systems (CARBBAS) freshwater dataset (with a total across all five datasets of 669 ASV-genome matches representing 580 unique genomes with inferred pH preferences; table S1 and fig. S5A). From the taxa with inferred pH preferences, the proportion of ASVs with available genome representatives was 10.6% in soils and 22.4% in freshwater on average (fig. S5A). Still, the taxonomic composition of the available genomes spanned 20 different phyla, with a general predominance of Actinobacteria and Proteobacteria among the matching genomes (fig. S5B), a result that is expected given that these two phyla are relatively well represented in genome databases (5).

We identified 332 gene types that had the same significant association with inferred pH preference in at least two datasets, while 56 of those genes had the same association with pH in at least three datasets across soil and freshwater habitats (Fig. 4 and table S2). We note that the taxonomic distinctiveness of the datasets and the generally even distribution of pH preferences across phyla support the notion that the identified associations between gene

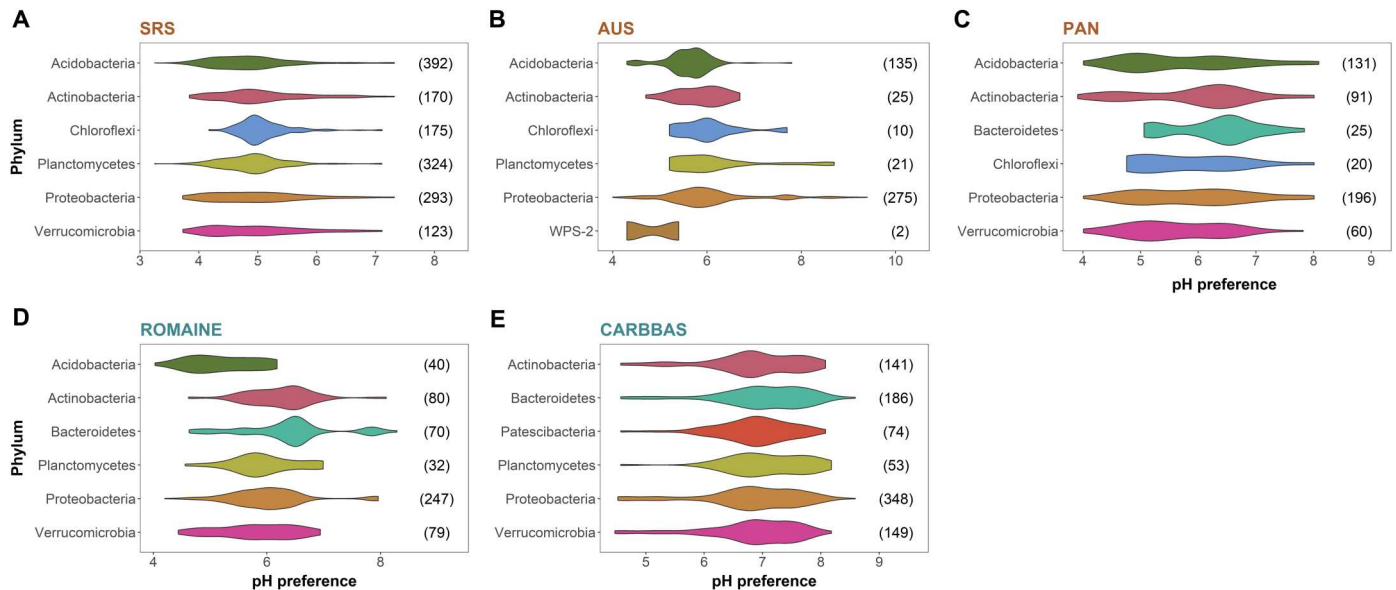


Fig. 2. Range of pH preferences of those amplicon sequence variants (ASVs) from which pH preferences could effectively be inferred from the biogeographical patterns in the marker gene sequencing data. The taxonomic assignments of the ASVs are shown at the phylum level with more complete taxonomic information included in fig. S2A. (A to E) correspond to the different datasets included in the study (brown labels, soil; turquoise labels, freshwater). Values at the end of the boxplots correspond to the number of ASVs with estimated pH preferences in each phylum. The length and width of the violins depict the entire range of pH preferences and the most frequent pH preferences of taxa within each phylum, respectively. SRS, Savannah River Site; AUS, Australia; PAN, Panama; ROMAINE, La Romaine watershed; CARBBAS, Carbon Biogeochemistry in Boreal Aquatic Systems.

types and inferred pH preferences were general and unlikely to be a product of the weak phylogenetic signal detected (figs. S3 and S5B; see also Materials and Methods for additional information). No gene types were identified as having a significant association with pH in all datasets, which is likely a result of pH adaptations not being conserved across taxa, these genes having other functionalities besides just pH adaptation, and the habitat-specific nature of some of the observed associations. The 56 gene types identified as having shared pH associations across both soil and freshwater habitats encoded for proteins known for their involvement in pH tolerance such as adenosine triphosphatases (ATPases), anion and cation transporters and antiporters, and alkaline and acidic phosphatases (Fig. 4). Generally, bacteria need to maintain pH homeostasis to maintain enzymatic function and cytoplasmic membrane stability, and they generally use four main mechanisms to cope with acid stress (41, 42). Across those genes that we identified as being associated with pH preferences (Fig. 4), we see evidence for all four mechanisms. First, proton-consuming reactions, notably decarboxylation and deamination of amino acids, buffer proton concentrations in the cytoplasm by incorporation of H^+ to metabolic by-products (43, 44). We identified genes for decarboxylases (*AAL_decaboxy*, soils), amino acid transporters (*AA_permease*, freshwater), carboxylate transporters (*TctA*, soils and freshwater), and amino acid deaminases (*Queuosine_synth*, soils and freshwater) associated with inferred pH preferences. Second, cells will produce basic compounds, such as ammonia released from urea to counter acidity. We found genes assigned to urea membrane transporters (*ureide_permeases*) overrepresented in taxa with low pH preference in soil and freshwater, as well as a gene for urease (*UreE_C*, soils) that hydrolyzes urea into ammonia (45). Third, bacteria can actively efflux protons to maintain intracellular pH levels. We identified

genes for a wide range of cation and anion efflux pumps such as the *Kdp* K^+ membrane transporters (*KdpACD*) that were overrepresented in taxa with low pH preference in all habitats. In contrast, Na^+/H^+ antiporters [*PhaGF*, *MnhG*, *MrpF*, and *YufB*; (33)] and anion transporters such as citrate (*CitMHS*, soils and freshwater) and lactate permeases (freshwater) were overrepresented in taxa with preferences for higher pH. Last, bacteria also modify the permeability of the cytoplasmic membrane and control the maturation and folding of proteins to limit acid stress. We identified genes for multiple hydrogenase quality control proteins (*HypCD*, *HycI*, and *HupF*) as overrepresented in taxa with low pH preference across soils and freshwater, with these genes known to be involved in acid stress response (45, 46). We also identified genes for process-specific proteins that act in a pH-dependent manner such as acidic phosphatases [Phosphoesterase and *CpsB_CapC* in soils and freshwater; (32)]. Malik *et al.* (28) analyzed genomic information from soil bacterial communities across a pH gradient of 4 to 8.5 and identified very similar genes and gene functions as found here (summarized in table S2). Overall, we found that 30 of the 56 gene types that we found are consistently associated with pH across habitats have previously been linked to bacterial pH adaptations in other studies (Fig. 4 and table S2).

While we cannot confirm that the genes that we identified are associated with pH preferences across multiple datasets actually represent specific bacterial adaptations to pH, our results make it possible to generate hypotheses about genes not previously involved in bacterial responses to pH. Our results expand on previous work by identifying multiple gene functions associated with pH preferences beyond transporter genes (14) while also confirming established associations between specific genes and bacterial pH adaptations (28). In addition, because most of our current knowledge regarding the

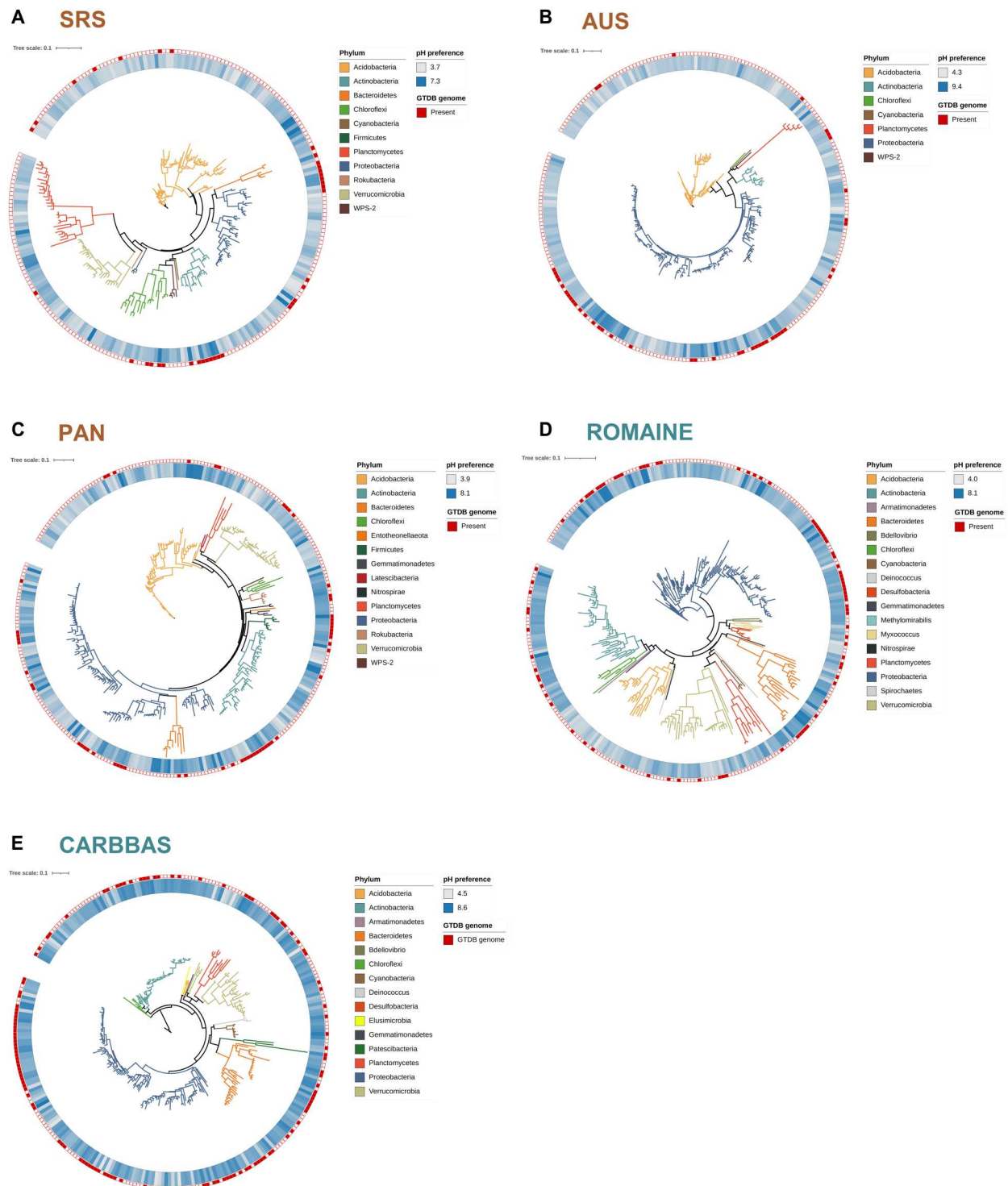


Fig. 3. Phylogenetic distributions of the most abundant amplicon sequence variants (ASVs) with estimated pH preferences in each dataset. Panels show the phylogenetic trees of the most abundant ASVs found in soil (A to C) and freshwater (D and E) habitats. Clades are colored by the affiliation at the phylum level. The inner ring depicts the estimated pH preference of the ASVs, and the outer ring indicates whether a genome with no more than one nucleotide mismatch to the amplicon sequence is available in the Genome Taxonomy Database (GTDG). The displays are maximum likelihood (ML) phylogenetic trees, and these trees only show the 300 most abundant ASVs per dataset, with the complete set of ASVs (numbers indicated in Fig. 2) used for the calculation of the phylogenetic signal in pH preferences (see "Phylogenetic visualization and analysis" section and fig. S3). The minimum pH preference is depicted in white, and the maximum is depicted in blue according to the legend. SRS, Savannah River Site; AUS, Australia; PAN, Panama; ROMAINE, La Romaine watershed; CARBBAS, Carbon Biogeochemistry in Boreal Aquatic Systems.

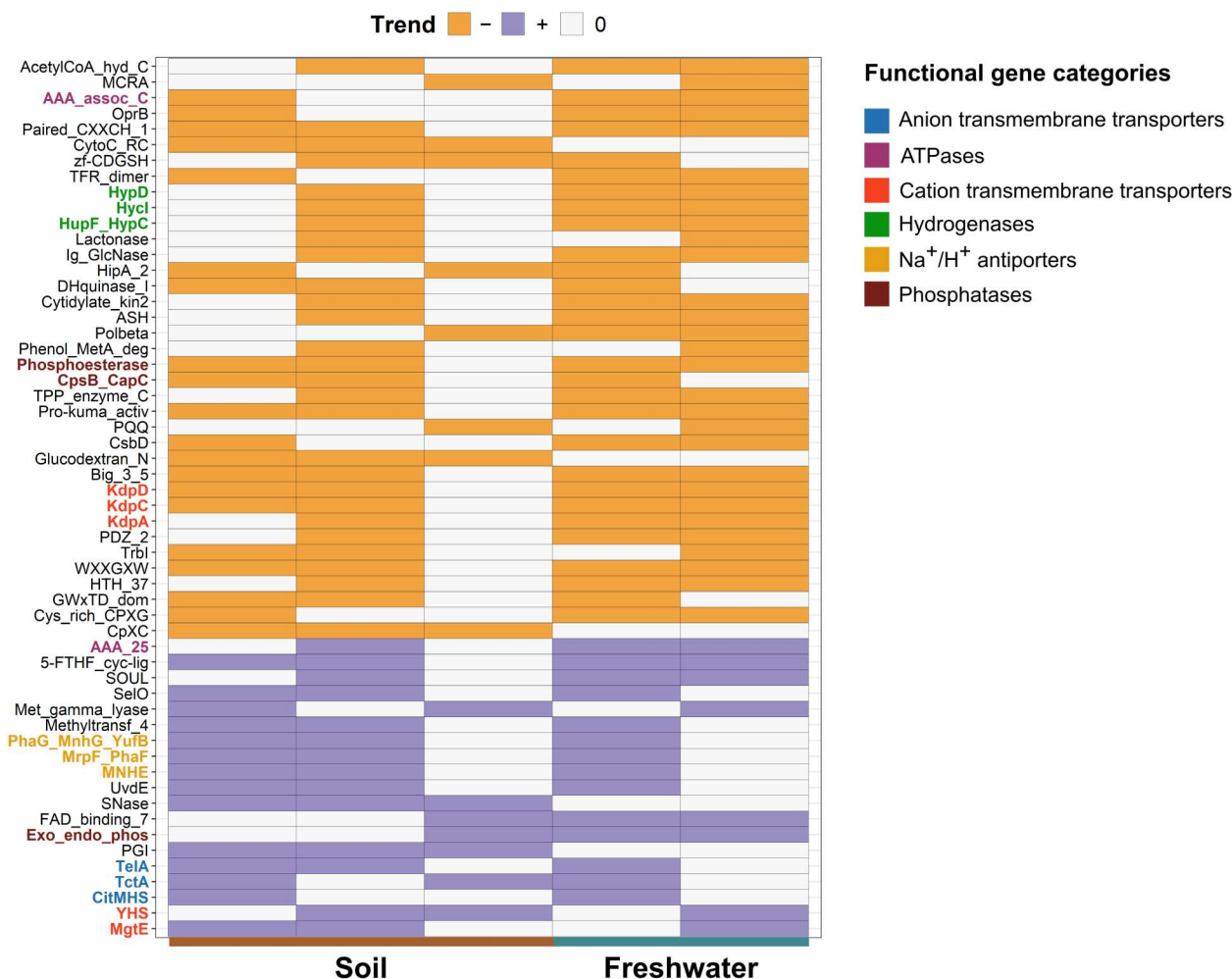


Fig. 4. Associations between gene categories and estimated pH preferences in bacteria across soils and freshwater systems. Positive trends (blue cells) indicate that the presence of the gene is associated with inferred bacterial preferences for higher pH, and negative trends (orange cells) indicate that presence of the gene is associated with bacterial preferences for lower pH conditions. We only included gene types that had consistent relationships to pH preference in at least three datasets and/or in both soil and freshwater. We highlighted categories with well-known involvement in pH tolerance and response in bacteria. For a complete summary of the putative functionality, Pfam identifier, and available references linking these specific gene types to pH adaptations see table S2. The x axis was sorted left to right in the order of Savannah River Site (SRS), Australia (AUS), Panama (PAN), La Romaine watershed (ROMAINE), and Carbon Biogeochemistry in Boreal Aquatic Systems (CARBBAS) datasets, respectively.

specific genes involved in pH tolerance is derived from in vitro studies focused on bacterial pathogens (47), our findings provide a basis to extend the investigation of bacterial adaptations to pH more broadly beyond those selected taxa.

Prediction of bacterial pH preference from genomic information alone

We incorporated the information on the presence/absence of the 56 gene types consistently associated with pH preference across habitats into a machine learning modeling framework to predict bacterial pH preferences. Preselecting gene types with consistent associations across habitats increased the likelihood that the model would maintain its predictive power on independent datasets (i.e., those containing genomes not included in this study). Across datasets, we obtained an average coefficient of determination (R^2) value of 0.80 for the linear regression between predicted and

observed pH preferences using the training data (Fig. 5). These genes encoded very diverse functions, from well-established ones such as transmembrane anion and cation transport, ATPase, and phosphatase activity, to functions less known to be involved in bacterial pH responses such as nucleases for DNA repair, endolysins, or the type V secretion system (table S2). With only presence/absence information of these 56 genes, we were able to predict the pH preferences of bacterial taxa that had not been used for model training. The validations conducted on a randomly selected subset of the genomes in each dataset (10% of genomes) had a mean absolute error (MAE) of 0.63 (MAE = 0.43 using the training data; Fig. 5), indicating that, with available presence/absence information of these 56 genes, this machine learning model can predict the pH preference of a given bacterial taxon with an accuracy of 0.63 pH units. Considering bacterial taxa generally have pH optima within 1 pH unit (48, 49), this error is relatively small. Likewise, the average

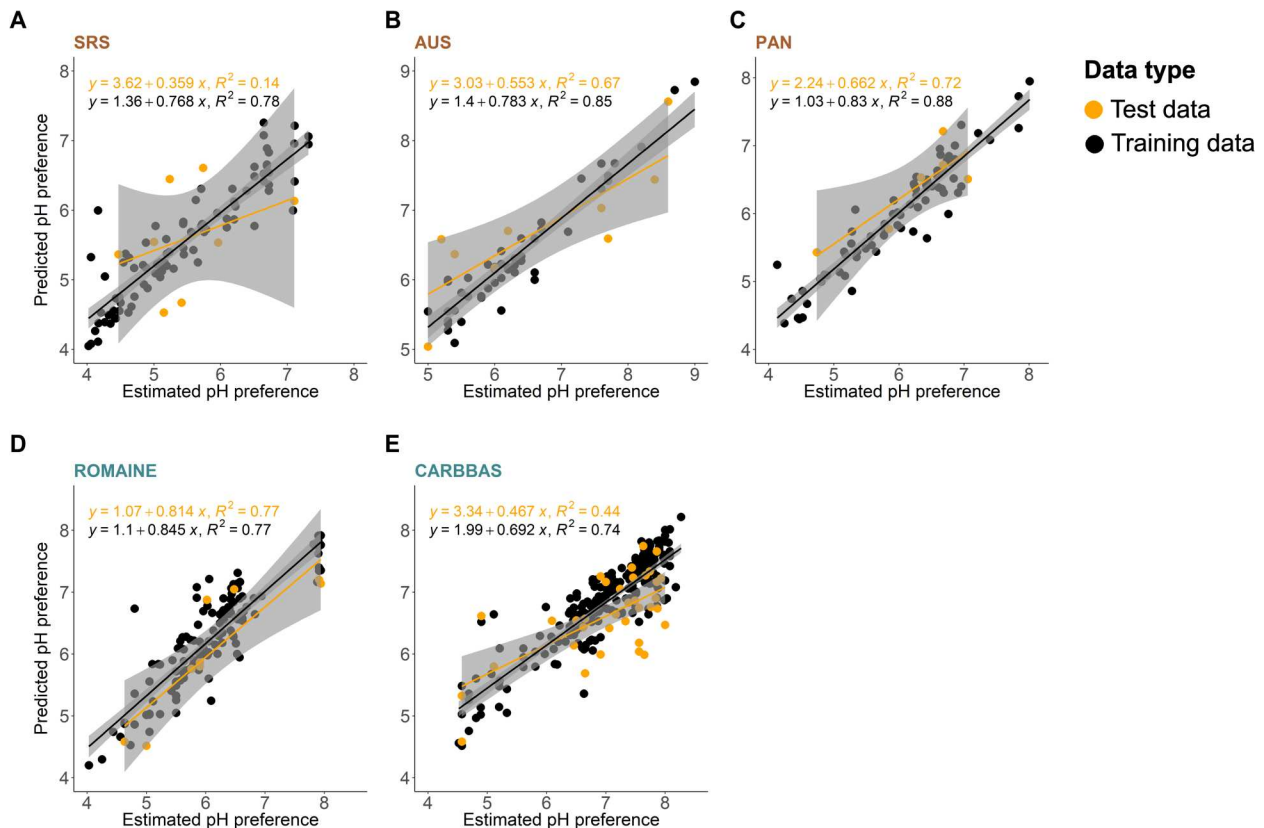


Fig. 5. Correlations between the estimated and predicted pH preferences of bacterial taxa based on a gradient boosted decision tree model built on the presence/absence of 56 top predictor gene types. (A to E) Each panel corresponds to a distinct dataset from either soil (brown) or freshwater (blue). The regressions on the plot include the predictions using the full dataset (black) and the independent validation using 10% of the genomes (orange) not included in the model development. The gray-shaded area depicts the SE around the fitted line. SRS, Savannah River Site; AUS, Australia; PAN, Panama; ROMAINE, La Romaine watershed; CARBBAS, Carbon Biogeochemistry in Boreal Aquatic Systems.

R^2 of the linear regression between the observed and predicted pH preferences in the independent validation set was 0.55. Note that, with this model, the absence of a gene is as informative for predictive purposes as gene presence (fig. S6) and that, due to the lack of data for taxa with estimated pH preferences below pH 4 and above pH 9, the model can only predict accurately within that range. We also note that our model predicted bacterial pH preferences across all five datasets, not necessarily pH optima for growth. Observed pH preferences correspond to the pH at which taxa achieve maximal abundance in nature, reflecting both the pH optimum for growth and the biotic and abiotic factors that may constrain bacterial abundances—the realized niche. We have made a detailed description of this model and how it can be applied to any genome of interest in <https://doi.org/10.6084/m9.figshare.22588963>.

We further validated our model using information on bacterial pH preferences estimated in a completely independent study of bacterial distributions in soils across the United Kingdom (38). The linear regression between the statistically estimated pH preference of taxa in the dataset and our model predictions had $R^2 = 0.21$ and MAE = 0.93 (fig. S7). Note that the estimated pH preferences in that study were obtained using a different approach from our study and that the pH preferences for many taxa were inferred using models with relatively weak fits, a point that the authors were careful to acknowledge (38). Despite the limitations of this

independent dataset and the limitations associated with our model, the correspondence between predicted and observed pH preferences from this independent study (and an independent set of genomes from our study; Fig. 5) supports the value of our approach and the model we developed. We emphasize the importance of including data from future studies that could measure actual pH preferences of bacterial isolates studied in vitro. While curated bacterial phenotypic information from culture collections can provide information on bacterial pH preferences, preexisting data on bacterial pH preferences of cultured isolates are now limited to the very narrow distribution of putative pH preferences [85.4% of pH preferences falling between 6 and 8 (5)], a pattern that most likely reflects the limited breadth of culturing conditions most commonly used in isolation efforts (14). Further quantification of microbial growth responses across large pH gradients under laboratory conditions coupled with whole genome sequencing is key to improving our knowledge of the genetic underpinnings of microbial adaptations to pH.

We show that biogeographical information can be combined with genomic information to infer and predict the pH preferences of bacterial taxa. This is important given the considerable effort often required to directly measure pH preferences of cultivated taxa in vitro and given that such in vitro assays are impossible for the majority of bacterial taxa that are resistant to cultivation. We

show that pH preferences can be inferred from genomic information alone, making it feasible to leverage the ever expanding genomic databases, including those which include MAGs and SAGs, to determine a key ecological attribute that cannot be readily determined from taxonomic or phylogenetic information alone (14). We not only identified genes that had been previously linked to pH tolerance via detailed studies of select bacteria but also identified genes that warrant further investigation as they have not previously been associated with pH adaptations.

Our approach demonstrates the feasibility of using genomic information to make predictions for other important traits that can be difficult to infer directly. In this sense, our work is similar to previous studies that have used genomic data to predict maximum potential growth rates (50, 51), oxygen tolerances (15), and temperature optima (12), among other traits. However, in those cases, data from cultivated isolates were used to develop the models linking genomic attributes to the trait values of interest. This represents an important bottleneck given that well-characterized cultivated isolates represent only a fraction of the phylogenetic and ecological diversity found in many environments. Instead, we demonstrate how biogeographical information, specifically distributions of taxa across environmental gradients of interest, combined with data from representative genomes, can be used to predict environmental preferences and identify the specific adaptations that may be associated with the ecological attributes. We expect that a similar approach could be used to quantify other relevant traits that have traditionally been difficult to infer directly for uncultivated taxa, including tolerances to changes in moisture, salinity, or heavy metals and other potentially toxic compounds.

The machine learning model presented here has the potential to aid the rational design of microbiomes where information on the pH preferences of bacterial taxa is needed. For example, several studies with N_2 -fixing rhizobia have successfully improved the symbiotic benefits to legume crops via the isolation and inoculation of acid resistant *Rhizobium* and *Bradyrhizobium* strains (52, 53). Forecasting invasive species spread can also benefit from genome-based predictions of pH (54–56), given the likely importance of pH as a factor limiting bacterial colonization of habitats. In addition, by predicting bacterial pH preferences, our model can aid the optimization of culturing conditions for any bacterial taxon with available genomic information. The coupling of biogeographical and genomic information can be successfully used to predict the environmental preferences of bacterial taxa, presenting opportunities for improving our trait-based understanding of microbial life.

MATERIALS AND METHODS

Datasets and sequence processing

We compiled five *16S rRNA* gene sequencing datasets that spanned broad gradients in pH and represented distinct ecosystem types. These datasets included two previously published soil datasets [soils collected from across Panama and across Australia; (18)], two previously published freshwater datasets [samples collected from streams, rivers, and lakes in Canada; (21, 22)], and one dataset of soils collected from the Savannah River Site (SRS) in South Carolina, United States.

For the SRS dataset, we collected soil samples in May 2021 from patches of savanna and surrounding forests of longleaf and loblolly pine that have been maintained since 2000 as part of the Corridor

Project experiment [e.g., (57)]. The SRS is a National Environmental Research Park. It is a U.S. Department of Energy site that is managed by the U.S. Department of Agriculture Forest Service. Each soil sample consisted of a homogenate of eight 5-cm-deep soil core subsamples. We extracted DNA using the DNeasy Power-Soil HTP 96 Kit (Qiagen) from well-mixed soil slurries consisting of 1 g of soil and 2 ml of deionized and autoclaved water. We amplified the V4 region of the *16S rRNA* gene using universal primers 515 forward and 806 reverse in duplicate polymerase chain reactions (PCRs). We cleaned and normalized amplicon samples using the SequelPrep Normalization Plate Kit (Applied Biosystems, Waltham, MA, USA) and sequenced a total of 240 soil samples, 22 extraction blanks (400 μ l of deionized and autoclaved water), and three PCR negative controls using paired-end Illumina MiSeq sequencing (300-cycle flow cell). For pH measurements, we created soil slurries consisting of 1 g of soil and 10 ml of deionized water, vortexed the slurries at maximum speed for 20 s, and then let them rest for about 1 hour.

The five datasets were selected as each one encompassed a broad range in measured sample pH values, each included a sufficiently large number of samples across the five pH gradients (>200 samples per dataset), and pH was strongly correlated with differences in overall bacterial community composition across each dataset (fig. S1). From all but the SRS dataset, we compiled the metadata from open-source databases and from the authors upon request and downloaded the DNA sequences from the Sequence Read Archive (SRA) of the National Center for Biotechnology Information. The raw sequences of the SRS project were deposited in the SRA under bioproject ID PRJNA898410. Additional details on these datasets are provided in fig. S1.

All datasets were analyzed using the same bioinformatic pipeline. Briefly, primers, adapters, and ambiguous bases were initially removed from the *16S rRNA* gene reads using cutadapt [v1.18; (58)]. Sequences were then quality-filtered, and ASVs were inferred using the DADA2 pipeline [v1.14.1; (59)]. Chimeric sequences were also removed, and taxonomic affiliations were determined against the SILVA SSU database [release 138; (60)]. The outputs were loaded into R (61) using the phyloseq package [v1.38.0; (62)] for downstream analyses.

Statistical inference of pH preferences

Singletons (ASVs represented by only a single read within a given dataset) were removed, and samples were rarefied to the minimum read number that ensured sufficient sequencing depth based on rarefaction curves (see table S1). As archaeal reads were relatively rare in most of these samples, we focused our analyses only on bacteria. Those bacterial ASVs that occurred in fewer than 20 samples per dataset were excluded from the analysis as ASVs needed to occur in a sufficiently large number of samples to effectively infer pH preferences (2 to 10% of ASVs passed this minimum threshold of occurrence per dataset; table S1). To infer pH preferences for individual ASVs in each dataset, we first generated 1000 randomized distributions of the relative abundance of each ASV across samples with replacement (i.e., each relative abundance value could be sampled more than once) and calculated the maximum value for each of these distributions. We then calculated 95% confidence intervals of these relative abundance maxima using the boot package (v1.3.28) in R. The extremes of these intervals in relative abundance maxima were then matched to the pH of the samples

where these ASVs achieved these relative abundance values, thus obtaining an estimated range of preferential pH for each ASV (i.e., the range of pH in which a given ASV consistently achieved maximal abundance across randomizations). All ASVs with inferred pH preferences that had ranges greater than 0.5 pH units were removed from downstream analyses as we were not confident that we could accurately infer a specific pH preference for those taxa. This stems from the assumption that we could only obtain confident statistical inferences for taxa with narrower pH preferences, excluding taxa with broader ranges in pH preferences to yield more accurate inferences. We visually inspected whether the abundance of each ASV in the nonrandomized dataset was within the identified preferred pH range to verify that our pH range filter was sufficiently stringent to exclude unreliable associations between ASVs and pH preferences. The midpoint of the pH range was taken as the estimated pH preference for each ASV. Our preliminary tests showed this bootstrapping approach to be more robust to zero inflation compared to alternative randomization approaches, generalized additive model, or logistic regression model fits that have previously been used for estimating environmental optima (34). We note that many factors, besides just pH, can influence the distributions of bacteria, and we restricted our analyses just to those taxa (ASVs) for which pH preferences could effectively be determined (4568 ASVs in total, 468 to 1614 ASVs per dataset; table S1).

Genome search and annotation

The ASVs with estimated pH preferences were matched to the annotated GTDB (release 207; (40)), using `vsearch` [v2.21.1; (63)] to identify representative genomes. We acknowledge that the representative genome identified for any given ASV will not necessarily be an identical match to the genome of that taxon in situ as even bacterial taxa with identical *16S rRNA* genes can have distinct genomes (64). However, this limitation poses a challenge to nearly all genomic analyses of bacteria from environmental samples, as even collections of MAGs would not necessarily capture the genomic heterogeneity that could exist at finer levels of taxonomic resolution. We also note that, when selecting representative genomes, we only allowed a single-base pair mismatch between the genomes and the *16S rRNA* gene amplicons, corresponding to a conservative 99.6% sequence similarity for the 250-base pair amplicons. In situations where a single ASV matched multiple genomes with the same similarity, we selected the most complete genome.

Predicted coding sequences for the ~62,000 unique bacterial genomes ("species clusters") available in the GTDB representative set were identified using `Prodigal` [v2.6.3; (65)]. The predicted coding sequences for each genome were aligned to the Pfam database [v35.0; (66)] using `HMMER` [v3; (67)] to obtain annotation of potential domains, genes, and gene families in each coding sequence. We discarded matches with a bit score lower than 10. This pipeline yielded a list of putative genes and domains found in each of the GTDB genomes that matched the ASVs identified from the samples included in this study with estimated pH preferences (580 genomes in total, 57 to 293 genomes per dataset; table S1). The copy numbers of genes and domains were binarized to presence/absence for further analyses.

Identification of genes associated with pH preference

We next determined associations between the estimated pH preferences of ASVs and the annotated genes in the corresponding representative genomes. We identified genes associated with pH preference by fitting generalized linear models with a binomial distribution and using the logit as a link function with core R functions. For each binarized gene function, we fitted a single model with the presence/absence of that gene type as the response variable and the estimated pH preference of each genome as the independent variable. We evaluated the statistical significance of the model coefficients using the Wald test as implemented with R core functions. We obtained the slope (positive or negative) of the relationship between pH preference and the presence/absence of each gene from the model estimates. We verified that the model estimates reliably reflected whether the relationship was positive or negative by plotting the proportion of genomes containing a particular gene type against the estimated pH preferences of those genomes. Last, we filtered out genes that had nonsignificant associations to pH preference in all datasets as well as those gene types that had significant associations in only one of the five datasets (as our goal was to identify consistent associations between genes and pH preferences across multiple datasets). We thus considered those gene types with the same significant association ($P < 0.05$ and same positive or negative direction of the model estimate) that occurred in two or more datasets to be associated with bacterial pH preferences. We verified that phylogenetic signal of pH preference among the representative genomes was very weak (in two datasets, Pearson's $r \sim 0.06$ to 0.07) or nonexistent (in three datasets) to justify not penalizing the generalized linear model estimates with phylogenetic information (fig. S8).

Phylogenetic visualization and analysis

We generated phylogenetic trees for each dataset that included only those ASVs for which pH preferences could be inferred. For each of those ASVs, we selected the corresponding highest quality full-length *16S rRNA* gene sequence from the SILVA SSU database [release 138; (60)], aligned those sequences with `MUSCLE` [v5; (68)], and created maximum likelihood (ML) trees using the RAxML approach with standard parameters [v8; (69)]. We visualized and edited the trees using `iTOL` [v6; (70)]. We additionally tested whether pH preference had a phylogenetic signal by calculating Blomberg's K (71) and Pagel's λ (72) with the depth at which pH preference had a phylogenetic signal estimated using phylogenetic correlograms (73). These three types of phylogenetic analyses were conducted using the R package `phylosignal` [v1.3; (74)] using ML trees constructed as described above with all ASVs from a given dataset for which pH preference could be inferred (table S1).

Prediction and independent validation of pH preferences from genomic information

We used the full set of genes associated with pH preference across habitats to train a gradient boosted decision tree model to predict bacterial pH preference from genomic information. The gene type presence/absence table was imported into Python, and pH preferences were predicted for all ASVs from a one-hot encoded gene matrix using gradient boosted decision trees created with the `XGBRegressor` function from Python's `xgboost` package [v1.6.2; (75)]. Hyperparameter optimization was implemented with the `hyperopt` package [v0.2.7; (76)] to select the best hyperparameters for

the XGBRegressor model. We validated the accuracy of the model by measuring its MAE on an independent 10% of the genomes in each dataset (i.e., the test set). We estimated the impact of each gene type on model predictions using Shapley additive explanations as integrated in Python's `xgboost` package. We also tested the model using a dataset of bacterial taxa with estimated pH preferences from soils across the United Kingdom (38), from which we selected taxa with unimodal relationships of abundance with pH and therefore more confident estimation of pH preference [models IV and V in (38)]. For each of these taxa, we obtained reference genomes using the same approach described for the ASVs in the other datasets and recorded the presence/absence of the predictive genes in those genomes to run the model.

Supplementary Materials

This PDF file includes:

Figs. S1 to S8

Tables S1 and S2

References

[View/request a protocol for this paper from Bio-protocol.](#)

REFERENCES AND NOTES

- M. Delgado-Baquerizo, A. M. Oliverio, T. E. Brewer, A. Benavent-González, D. J. Eldridge, R. D. Bardgett, F. T. Maestre, B. K. Singh, N. Fierer, A global atlas of the dominant bacteria found in soil. *Science* **359**, 320–325 (2018).
- D. C. Laughlin, Applying trait-based models to achieve functional targets for theory-driven ecological restoration. *Ecol. Lett.* **17**, 771–784 (2014).
- M. G. Mathipa, M. S. Thantsha, Probiotic engineering: Towards development of robust probiotic strains with enhanced functional properties and for targeted control of enteric pathogens. *Gut Pathog.* **9**, 1–17 (2017).
- S. J. Green, C. B. Brookson, N. A. Hardy, L. B. Crowder, Trait-based approaches to global change ecology: Moving from description to prediction. *Proc. Biol. Sci.* **289**, 20220071 (2022).
- J. S. Madin, D. A. Nielsen, M. Brbic, R. Corkrey, D. Danko, K. Edwards, M. K. M. Engqvist, N. Fierer, J. L. Geoghegan, M. Gillings, N. C. Kyrpides, E. Litchman, C. E. Mason, L. Moore, S. L. Nielsen, I. T. Paulsen, N. D. Price, T. B. K. Reddy, M. A. Richards, E. P. C. Rocha, T. M. Schmidt, H. Shaaban, M. Shukla, F. Supek, S. G. Tetu, S. Vieira-Silva, A. R. Wattam, D. A. Westfall, M. Westoby, A synthesis of bacterial and archaeal phenotypic trait data. *Sci. Data* **7**, 170 (2020).
- A. Cébron, E. Zeghal, P. Usseglio-Polatera, A. Meyer, P. Bauda, F. Lemmel, C. Leyval, F. Maunoury-Danger, BactoTraits – A functional trait database to evaluate how natural and man-induced changes influence the assembly of bacterial communities. *Ecol. Indic.* **130**, 108047 (2021).
- A. D. Steen, A. Crits-Christoph, P. Carini, K. M. DeAngelis, N. Fierer, K. G. Lloyd, J. Cameron Thrash, High proportions of bacteria and archaea across most biomes remain uncultured. *ISME J.* **13**, 3126–3130 (2019).
- E. Bååth, E. Kritzberg, pH tolerance in freshwater bacterioplankton: Trait variation of the community as measured by leucine incorporation. *Appl. Environ. Microbiol.* **81**, 7411–7419 (2015).
- K. K. Treseder, S. N. Klivlin, C. V. Hawkes, Evolutionary trade-offs among decomposers determine responses to nitrogen enrichment. *Ecol. Lett.* **14**, 933–938 (2011).
- N. Lustenhouwer, D. S. Maynard, M. A. Bradford, D. L. Lindner, B. Oberle, A. E. Zanne, T. W. Crowther, A trait-based understanding of wood decomposition by fungi. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 11551–11558 (2020).
- J. T. Lennon, Z. T. Aanderud, B. K. Lehmkuhl, D. R. Schoolmaster Jr., Mapping the niche space of soil microorganisms using taxonomy and traits. *Ecology* **93**, 1867–1879 (2012).
- D. B. Sauer, D.-N. Wang, Predicting the optimal growth temperatures of prokaryotes using only genome derived features. *Bioinformatics* **35**, 3224–3231 (2019).
- P.-A. Chaumel, A. J. Mussig, P. Hugenholtz, D. H. Parks, GTDB-Tk: A toolkit to classify genomes with the Genome Taxonomy Database. *Bioinformatics* **36**, 1925–1927 (2020).
- A. Barberán, H. Caceres Velazquez, S. Jones, N. Fierer, Hiding in plain sight: Mining bacterial species records for phenotypic trait information. *mSphere* **2**, e00237-17 (2017).
- J. Jabłońska, D. S. Tawfik, The number and type of oxygen-utilizing enzymes indicates aerobic vs. anaerobic phenotype. *Free Radic. Biol. Med.* **140**, 84–92 (2019).
- A. Ö. C. Dupont, R. I. Griffiths, T. Bell, D. Bass, Differences in soil micro-eukaryotic communities over soil pH gradients are strongly driven by parasites and saprotrophs. *Environ. Microbiol.* **18**, 2010–2024 (2016).
- J. Davison, M. Moora, M. Semchenko, S. B. Adenan, T. Ahmed, A. A. Akhmetzhanova, J. M. Alatalo, S. Al-Quraishy, E. Andriyanova, S. Anslan, M. Bahram, A. Batbaatar, C. Brown, C. G. Bueno, J. Cahill, J. J. Cantero, B. B. Casper, M. Cherosov, S. Chideh, A. P. Coelho, M. Coghill, G. Decocq, S. Dudov, E. C. Fabiano, V. E. Fedosov, L. Fraser, S. I. Glassman, A. Helm, H. A. L. Henry, B. Héroult, I. Hiiesalu, I. Hiiesalu, W. N. Hozzein, P. Kohout, U. Kõljalg, K. Koorem, L. Laanisto, Ü. Mander, L. Mucina, J.-P. Munyampundu, L. Nukenkamp, Ü. Niinemets, C. Nyamukondiwa, J. Oja, V. Onipchenko, M. Pärtel, C. Phosri, S. Pölme, K. Püssa, A. Ronk, A. Saitta, O. Semboli, S.-K. Sepp, A. Seregín, S. Sudheer, C. P. Peña-Venegas, C. Paz, T. Vahter, M. Vasar, A. J. Veraart, L. Tedersoo, M. Zobel, M. Öpik, Temperature and pH define the realised niche space of arbuscular mycorrhizal fungi. *New Phytol.* **231**, 763–776 (2021).
- A. M. Oliverio, A. Bissett, K. McGuire, K. Saltonstall, B. L. Turner, N. Fierer, The role of phosphorus limitation in shaping soil bacterial communities and their metabolic capabilities. *mBio* **11**, e01718–e01720 (2020).
- N. Fierer, R. B. Jackson, The diversity and biogeography of soil bacterial communities. *Proc. Natl. Acad. Sci. U.S.A.* **103**, 626–631 (2006).
- R. Ortiz-Álvarez, J. Cáliz, L. Camarero, E. O. Casamayor, Regional community assembly drivers and microbial environmental sources shaping bacterioplankton in an alpine lacustrine district (Pyrenees, Spain). *Environ. Microbiol.* **22**, 297–309 (2020).
- J. P. Niño-García, C. Ruiz-González, P. A. Del Giorgio, Interactions between hydrology and water chemistry shape bacterioplankton biogeography across boreal freshwater networks. *ISME J.* **10**, 1755–1766 (2016).
- M. Stadler, P. A. del Giorgio, Terrestrial connectivity, upstream aquatic history and seasonality shape bacterial community assembly within a large boreal aquatic network. *ISME J.* **16**, 937–947 (2022).
- J. F. Power, C. R. Carere, C. K. Lee, G. L. J. Wakerley, D. W. Evans, M. Button, D. White, M. D. Climo, A. M. Hinze, X. C. Morgan, I. R. McDonald, S. C. Cary, M. B. Stott, Microbial biogeography of 925 geothermal springs in New Zealand. *Nat. Commun.* **9**, 2876 (2018).
- Q. Jin, M. F. Kirk, pH as a primary control in environmental microbiology: 1. Thermodynamic perspective. *Front. Environ. Sci.* **6**, 21 (2018).
- M. A. Oberhardt, R. Zarecki, S. Gronow, E. Lang, H.-P. Klenk, U. Gophna, E. Ruppín, Harnessing the landscape of microbial culture media to predict new organism–media pairings. *Nat. Commun.* **6**, 8493 (2015).
- W. H. Hartman, C. J. Richardson, R. Vilgalys, G. L. Bruland, Environmental and anthropogenic controls over bacterial communities in wetland soils. *Proc. Natl. Acad. Sci. U.S.A.* **105**, 17842–17847 (2008).
- C. L. Lauber, M. Hamady, R. Knight, N. Fierer, Pyrosequencing-based assessment of soil pH as a predictor of soil bacterial community structure at the continental scale. *Appl. Environ. Microbiol.* **75**, 5111–5120 (2009).
- A. A. Malik, B. C. Thomson, A. S. Whiteley, M. Bailey, R. I. Griffiths, Bacterial physiological adaptations to contrasting edaphic conditions identified using landscape scale metagenomics. *mBio* **8**, 799–816 (2017).
- R. T. Jones, M. S. Robeson, C. L. Lauber, M. Hamady, R. Knight, N. Fierer, A comprehensive survey of soil acidobacterial diversity using pyrosequencing and clone library analyses. *ISME J.* **3**, 442–453 (2009).
- J. Rousk, E. Bååth, P. C. Brookes, C. L. Lauber, C. Lozupone, J. G. Caporaso, R. Knight, N. Fierer, Soil bacterial and fungal communities across a pH gradient in an arable soil. *ISME J.* **4**, 1340–1351 (2010).
- N. Guan, L. Liu, Microbial response to acid stress: Mechanisms and applications. *Appl. Microbiol. Biotechnol.* **104**, 51–65 (2020).
- Y. Derman, H. Söderholm, M. Lindström, H. Korkeala, Role of *csp* genes in NaCl, pH, and ethanol stress response and motility in *Clostridium botulinum* ATCC 3502. *Food Microbiol.* **46**, 463–470 (2015).
- M. Ito, M. Morino, T. A. Krulwich, Mrp antiporters have important roles in diverse bacteria and archaea. *Front. Microbiol.* **8**, 2325 (2017).
- J. C. Benavides, D. H. Vitt, Response curves and the environmental limits for peat-forming species in the northern Andes. *Plant Ecol.* **215**, 937–952 (2014).
- N. Fierer, J. L. Morse, S. T. Berthrong, E. S. Bernhardt, R. B. Jackson, Environmental controls on the landscape-scale biogeography of stream bacterial communities. *Ecology* **88**, 2162–2173 (2007).
- A. M. Kielak, C. C. Barreto, G. A. Kowalchuk, J. A. van Veen, E. E. Kuramae, The ecology of Acidobacteria: Moving beyond genes and genomes. *Front. Microbiol.* **7**, 744 (2016).

37. J. A. F. Diniz-Filho, T. Santos, T. F. Rangel, L. M. Bini, A comparison of metrics for estimating phylogenetic signal under alternative evolutionary models. *Genet. Mol. Biol.* **35**, 673–679 (2012).
38. B. Jones, T. Goodall, P. B. L. George, H. S. Gweon, J. Puissant, D. S. Read, B. A. Emmett, D. A. Robinson, D. L. Jones, R. I. Griffiths, Beyond taxonomic identification: Integration of ecological responses to a soil bacterial 16S rRNA gene database. *Front. Microbiol.* **12**, 682886 (2021).
39. S. Mukherjee, R. Seshadri, N. J. Varghese, E. A. Eloë-Fadros, J. P. Meier-Kolthoff, M. Göker, R. C. Coates, M. Hadjithomas, G. A. Pavlopoulos, D. Paez-Espino, Y. Yoshikuni, A. Visel, W. B. Whitman, G. M. Garrity, J. A. Eisen, P. Hugenholtz, A. Pati, N. N. Ivanova, T. Woyke, H.-P. Klenk, N. C. Kyrpides, 1,003 reference genomes of bacterial and archaeal isolates expand coverage of the tree of life. *Nat. Biotechnol.* **35**, 676–683 (2017).
40. D. H. Parks, M. Chuvochina, C. Rinke, A. J. Mussig, P.-A. Chaumeil, P. Hugenholtz, GTDB: An ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy. *Nucleic Acids Res.* **50**, D785–D794 (2022).
41. P. A. Lund, D. De Biase, O. Liran, O. Scheler, N. P. Mira, Z. Cetecioglu, E. N. Fernández, S. Bover-Cid, R. Hall, M. Sauer, C. O'Byrne, Understanding how microorganisms respond to acid pH is central to their control and successful exploitation. *Front. Microbiol.* **11**, 556140 (2020).
42. P. D. Cotter, C. Hill, Surviving the acid test: Responses of gram-positive bacteria to low pH. *Microbiol. Mol. Biol. Rev.* **67**, 429–453 (2003).
43. T. A. Krulwich, G. Sachs, E. Pagan, Molecular aspects of bacterial pH sensing and homeostasis. *Nat. Rev. Microbiol.* **9**, 330–343 (2011).
44. H. A. Danhof, S. Vylkova, E. M. Vesely, A. E. Ford, M. Gonzalez-Garay, M. C. Lorenz, Robust extracellular pH modulation by *Candida albicans* during growth in carboxylic acids. *mBio* **7**, e01646-16 (2016).
45. Y. Wen, E. A. Marcus, U. Matrubutham, M. A. Gleeson, D. R. Scott, G. Sachs, Acid-adaptive genes of *Helicobacter pylori*. *Infect. Immun.* **71**, 5921–5939 (2003).
46. E. T. Hayes, J. C. Wilks, P. Sanfilippo, E. Yohannes, D. P. Tate, B. D. Jones, M. D. Radmacher, S. S. BonDurant, J. L. Slonczewski, Oxygen limitation modulates pH regulation of catabolism and hydrogenases, multidrug transporters, and envelope composition in *Escherichia coli* K-12. *BMC Microbiol.* **6**, 89 (2006).
47. P. Lund, A. Tramonti, D. De Biase, Coping with low pH: Molecular strategies in neutrophilic bacteria. *FEMS Microbiol. Rev.* **38**, 1091–1125 (2014).
48. J. B. Russell, D. B. Dombrowski, Effect of pH on the efficiency of growth by pure cultures of rumen bacteria in continuous culture. *Appl. Environ. Microbiol.* **39**, 604–610 (1980).
49. J. J. Therion, A. Kistner, J. H. Kornelius, Effect of pH on growth rates of rumen amylolytic and lactilytic bacteria. *Appl. Environ. Microbiol.* **44**, 428–434 (1982).
50. S. Vieira-Silva, E. P. C. Rocha, The systemic imprint of growth and its uses in ecological (meta)genomics. *PLoS Genet.* **6**, e1000808 (2010).
51. J. L. Weissman, S. Hou, J. A. Fuhrman, Estimating maximal microbial growth rates from cultures, metagenomes, and single cells via codon usage patterns. *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2016810118 (2021).
52. A. Chinnaswamy, B. Dhar, Symbiotic effectiveness of acid-tolerant *Bradyrhizobium* strains with soybean in low pH soil. *African J. Biotechnol.* **5**, 842–845 (2006).
53. D. P. Oliveira, M. A. de Figueiredo, B. L. Soares, O. H. Stivanin Teixeira, F. A. Dias Martins, M. Rufini, C. Peixoto Chain, R. Pereira Reis, A. Ramalho de Moraes, F. M. de Souza Moreira, M. J. B. de Andrade, Acid tolerant *Rhizobium* strains contribute to increasing the yield and profitability of common bean in tropical soils. *J. Soil Sci. Plant Nutr.* **17**, 922–933 (2017).
54. J. F. Briand, C. Le Boulanger, J. F. Humbert, C. Bernard, P. Dufour, *Cylindrospermopsis raciborskii* (cyanobacteria) invasion at mid-latitudes: Selection, wide physiological tolerance, orglobalwarming? *J. Phycol.* **40**, 231–238 (2004).
55. S. Pertola, H. Kuosa, R. Olsonen, Is the invasion of *Prorocentrum minimum* (Dinophyceae) related to the nitrogen enrichment of the Baltic Sea? *Harmful Algae* **4**, 481–492 (2005).
56. P. C. Reid, D. G. Johns, M. Edwards, M. I. Starr, M. Poulin, P. Snoeijs, A biological consequence of reducing Arctic ice cover: Arrival of the Pacific diatom *Neodenticula seminae* in the North Atlantic for the first time in 800000 years. *Glob. Chang. Biol.* **13**, 1910–1921 (2007).
57. J. J. Tewksbury, D. J. Levey, N. M. Haddad, S. Sargent, J. L. Orrock, A. Weldon, B. J. Danielson, J. Brinkerhoff, E. I. Damschen, P. Townsend, Corridors affect plants, animals, and their interactions in fragmented landscapes. *Proc. Natl. Acad. Sci. U.S.A.* **99**, 12923–12926 (2002).
58. M. Martin, Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnetjournal* **17**, 10–12 (2011).
59. B. J. Callahan, P. J. McMurdie, S. P. Holmes, Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. *ISME J.* **11**, 2639–2643 (2017).
60. C. Quast, E. Pruesse, P. Yilmaz, J. Gerken, T. Schweer, P. Yarza, J. Peplies, F. O. Glöckner, The SILVA ribosomal RNA gene database project: Improved data processing and web-based tools. *Nucleic Acids Res.* **41**, D590–D596 (2013).
61. R: The R Project for Statistical Computing; www.r-project.org/.
62. P. J. McMurdie, S. Holmes, phyloseq: An R package for reproducible interactive analysis and graphics of microbiome census data. *PLoS ONE* **8**, e61217 (2013).
63. T. Rognes, T. Flouri, B. Nichols, C. Quince, F. Mahé, VSEARCH: A versatile open source tool for metagenomics. *PeerJ.* **4**, e2584 (2016).
64. H. B. Hassler, B. Probert, C. Moore, E. Lawson, R. W. Jackson, B. T. Russell, V. P. Richards, Phylogenies of the 16S rRNA gene and its hypervariable regions lack concordance with core genome phylogenies. *Microbiome* **10**, 104 (2022).
65. D. Hyatt, G. L. Chen, P. F. LoCascio, M. L. Land, F. W. Larimer, L. J. Hauser, Prodigal: Prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11**, 119 (2010).
66. J. Mistry, S. Chuguransky, L. Williams, M. Qureshi, G. A. Salazar, E. L. L. Sonnhammer, S. C. E. Tosatto, L. Paladini, S. Raj, L. J. Richardson, R. D. Finn, A. Bateman, Pfam: The protein families database in 2021. *Nucleic Acids Res.* **49**, D412–D419 (2021).
67. R. D. Finn, J. Clements, S. R. Eddy, HMMER web server: Interactive sequence similarity searching. *Nucleic Acids Res.* **39**, W29–W37 (2011).
68. R. C. Edgar, MUSCLE: A multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* **5**, 113 (2004).
69. A. Stamatakis, RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).
70. I. Letunic, P. Bork, Interactive Tree Of Life (iTOL) v5: An online tool for phylogenetic tree display and annotation. *Nucleic Acids Res.* **49**, W293–W296 (2021).
71. S. P. Blomberg, T. Garland Jr., A. R. Ives, Testing for phylogenetic signal in comparative data: Behavioral traits are more labile. *Evolution* **57**, 717–745 (2003).
72. M. Pagel, Inferring the historical patterns of biological evolution. *Nature* **401**, 877–884 (1999).
73. J. L. Gittleman, M. Kot, Adaptation: Statistics and a null model for estimating phylogenetic effects. *Syst. Biol.* **39**, 227–241 (1990).
74. F. Keck, F. Rimet, A. Bouchez, A. Franc, phylosignal: An R package to measure, test, and explore the phylogenetic signal. *Ecol. Evol.* **6**, 2774–2780 (2016).
75. T. Chen, C. Guestrin, XGBoost: A scalable tree boosting system, in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Association for Computing Machinery, 2016), 785–794.
76. J. Bergstra, D. Yamins, D. Cox, Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures. **28**, 115–123 (2013).
77. J. Snider, I. Gutschke, M. Lin, S. Baby, B. Cox, G. Butland, J. Greenblatt, A. Emili, W. A. Houry, Formation of a distinctive complex between the inducible bacterial lysine decarboxylase and a novel AAA+ ATPase. *J. Biol. Chem.* **281**, 1532–1546 (2006).
78. A. Alam, J. E. Bröms, R. Kumar, A. Sjöstedt, The role of ClpB in bacterial stress responses and virulence. *Front. Mol. Biosci.* **8**, 668910 (2021).
79. J. J. Lensbouwer, R. P. Doyle, Secondary transport of metal–citrate complexes: The CitMHS family. *Crit. Rev. Biochem. Mol. Biol.* **45**, 453–462 (2010).
80. Z.-C. Yuan, P. Liu, P. Saenkham, K. Kerr, E. W. Nester, Transcriptome profiling and functional analysis of *Agrobacterium tumefaciens* reveals a general conserved response to acidic conditions (pH 5.5) and a complex acid-mediated signaling involved in *Agrobacterium*-plant interactions. *J. Bacteriol.* **190**, 494–507 (2008).
81. E. Khalikova, P. Susi, T. Korpela, Microbial dextran-hydrolyzing enzymes: Fundamentals and applications. *Microbiol. Mol. Biol. Rev.* **69**, 306–325 (2005).
82. W.-J. Jung, J.-H. Kuk, K.-Y. Kim, K.-C. Jung, R.-D. Park, Purification and characterization of exo-β-d-glucosaminidase from *Aspergillus fumigatus* S-26. *Protein Expr. Purif.* **45**, 125–131 (2006).
83. M. C. Cholo, M. T. Matjokotja, A. G. Osman, R. Anderson, Role of the *kdpDE* regulatory operon of *Mycobacterium tuberculosis* in modulating bacterial growth in vitro. *Front. Genet.* **12**, 698875 (2021).
84. A. Ballal, B. Basu, S. K. Apte, The Kdp-ATPase system and its regulation. *J. Biosci.* **32**, 559–568 (2007).
85. H. Asha, J. Gowrishankar, Regulation of kdp operon expression in *Escherichia coli*: Evidence against turgor as signal for transcriptional control. *J. Bacteriol.* **175**, 4528–4537 (1993).
86. T. Hiramatsu, K. Kodama, T. Kuroda, T. Mizushima, T. Tsuchiya, A putative multisubunit Na⁺/H⁺ antiporter from *Staphylococcus aureus*. *J. Bacteriol.* **180**, 6642–6648 (1998).
87. P. Putnok, A. Kereszt, T. Nakamura, G. Endre, E. Grosskopf, P. Kiss, Á. Kondorosi, The pha gene cluster of *Rhizobium meliloti* involved in pH adaptation and symbiosis encodes a novel type of K⁺ efflux system. *Mol. Microbiol.* **28**, 1091–1101 (1998).
88. M. Ito, A. A. Guffanti, B. Oudega, T. A. Krulwich, Mrp, a multigene, multifunctional locus in *Bacillus subtilis* with roles in resistance to cholate and to Na⁺ and in pH homeostasis. *J. Bacteriol.* **181**, 2394–2402 (1999).
89. J. M. Chirgwin, T. F. Parsons, E. A. Noltmann, Mechanistic implications of the pH independence of inhibition of phosphoglucose isomerase by neutral sugar phosphates. *J. Biol. Chem.* **250**, 7277–7279 (1975).

90. N. S. DeSilva, P. A. Quinn, Characterization of phospholipase A1, A2, C activity in Ureaplasma urealyticum membranes. *Mol. Cell. Biochem.* **201**, 159–167 (1999).
91. M. B. Marques, P. F. Weller, A. Nicholson-Weller, Growth in acidic media increases production of phosphatidylinositol-specific phospholipase C by *Staphylococcus aureus*. *Curr. Microbiol.* **25**, 125–128 (1992).
92. M. Comellas-Bigler, K. Maskos, R. Huber, H. Oyama, K. Oda, W. Bode, 1.2 Å crystal structure of the serine carboxyl proteinase pro-kumamolisin; structure of an intact pro-subtilase. *Structure* **12**, 1313–1323 (2004).
93. M. Caza, J. W. Kronstad, Shared and distinct mechanisms of iron acquisition by bacterial and fungal pathogens of humans. *Front. Cell. Infect. Microbiol.* **3**, 80 (2013).

Acknowledgments: We thank all the people involved in the acquisition of the data compiled in this study. **Funding:** This work was supported by the Swiss National Science Foundation (Early Postdoc Mobility P2EZP3_199849 to J.R.). U.S. National Science Foundation (Graduate Research Fellowship Program DGE 1650115 to C.C.W.; OPP 2133684 and AW5809-826664 to E.S.-O., N.F., and M.H.), NSERC/Hydro-Québec Industrial Research Chair in Carbon Biogeochemistry in Boreal Aquatic Systems (387312-14, IRC 2015-2019 to M.S.; 387312-07, IRC 2009-2014 to J.P.N.-G.),

National Science Foundation (awards 1912729 and 1913501), and U.S. Department of Energy to the U.S. Department of Agriculture-Forest Service-Savannah River under interagency agreement 89303720SEM000037. **Author contributions:** Study design: J.R., M.H., and N.F. Data provision: C.C.W., M.S., J.P.N.-G., and N.F. Bioinformatics: M.H. Machine learning model: E.S.-O. Statistical analyses: J.R. Visualization: J.R. Writing—original draft: J.R. and N.F. Writing—review and editing: J.R., E.S.-O., M.H., C.C.W., M.S., J.P.N.-G., and N.F. **Competing interests:** The authors declare that they have no competing interests. **Data and materials availability:** All data needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials. Raw sequencing data from the SRS dataset have been uploaded to the SRA (bioproject ID PRJNA898410), and other data are publicly available from Figshare (<https://doi.org/10.6084/m9.figshare.22310341> and <https://doi.org/10.6084/m9.figshare.22310383>).

Submitted 18 November 2022

Accepted 27 March 2023

Published 28 April 2023

10.1126/sciadv.adf8998